BioHealthBase
BioDefense Public Health Database

BRC4 Meeting
December 7, 2006
University of Alabama - Birmingham

BioHealthBase Yesterday, Today, and Tomorrow

The BioHealthBase Bioinformatics Resource Center for Biodefense and Emerging/Re-emerging Infectious Disease

Kevin J. Biersack
Northrop Grumman Health Solutions
Rockville, MD

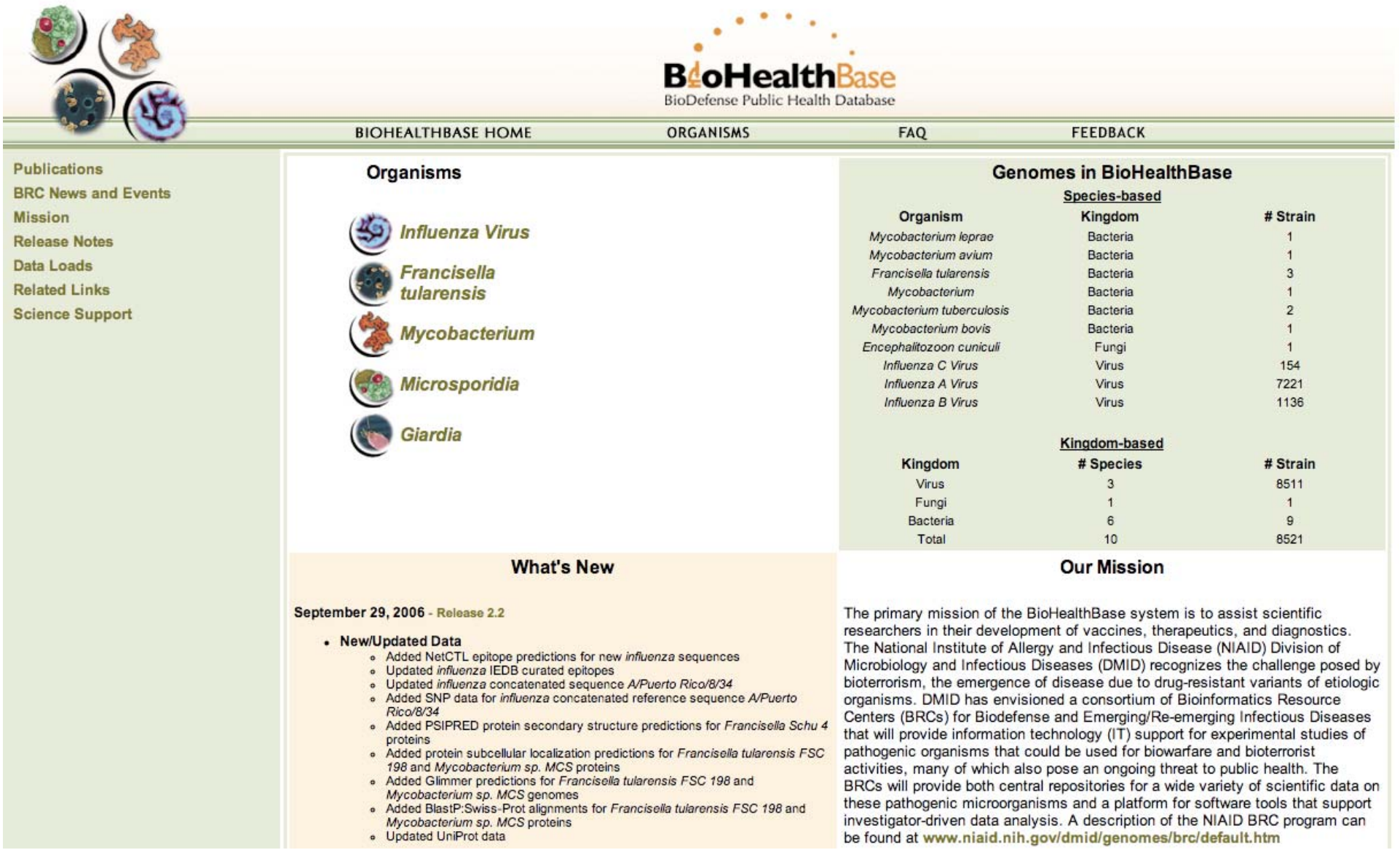
UT SOUTHWESTERN MEDICAL CENTER

NATIONAL INSTITUTES OF HEALTH

NIAID

NORTHROP GRUMMAN
Information Technology

BioHealthBase: Home Page



www.biohealthbase.org

V1.0 Baseline Release – S/W

- Database Search and Display
  - Virus (*Influenza*)
    - Gene, Epitope, Public Database Identifier, Polymorphism
  - Bacteria (*MTB and Tularemia*)
    - Gene Search, Public Database Identifier, Generic locus, Generic GO
  - Retrieve the following detailed gene and protein information for a selected gene:
    - Gene Identification, Gene and feature location, Operon identification (Bacteria Only), Protein identification, Protein domain and motif mappings, Preliminary Protein Localization (Bacteria), Preliminary epitope hits for MHC superfamilies (Influenza), GO Classification, Database cross-reference, Literature References, Data sources (with footnotes)
  - Genomic browsing via the GMOD GBrowse tool (bacteria only)
  - Data download:
    - Genome, gene, and feature sequences in FASTA format
    - Annotation in GFF3 format
- Data Processing (Pre-computes)
  - Semi-automated data processing pipeline (initially, BLAST using the NCBI non-redundant amino acid (NRAA) database)
  - Creation of consensus sequences for each influenza serological type
  - Influenza frequency of minor allele computation with comparison to the consensus influenza sequences for each serological type
  - Initial protein localization computations for bacterial organisms
  - Preliminary MHC class I epitope prediction for influenza

V1.0 Baseline Release - Data

- *Francisella tularensis* genome sequence and annotation
  - *Schu S4* - Provisional RefSeq
- *Mycobacterium tuberculosis* genome sequence and annotation
  - *Mycobacterium tuberculosis* - CDC 1551 - Provisional RefSeq
  - *Mycobacterium tuberculosis* - H37Rv - Provisional RefSeq
  - *M. bovis*, *M. leprae*, and *M. avium* - Provisional RefSeqs
- *Influenza virus* (types A, B, and C) - from NCBI
  - All influenza nucleotide and protein sequences provided to NCBI
- Protein information, integrated with corresponding NCBI annotation and sequences:
  - UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, and Pfam (domains and motifs)
- Generated Data:
  - Protein localization (PSortB, SubLoc, SignalP, DAS - bacteria only)
  - MHC epitopes (NetCTL - *influenza* only)
  - Sequence polymorphism analysis etc. (MUSCLE plus custom software - *influenza* only)
- Pathway Data:
  - BioCyc metabolic pathways (bacteria only)
  - Operons (BioCyc Pathway Tools - bacteria only)
  - Reactome interferon pathway (*influenza* only)

Through V2.2 Release – S/W

- Database Search and Display
  - *Encephalitozoon cuniculi* (Microsporidia)
    - Gene, Locus, Public Database Identifier, GO, and Pfam Domain and Motif
  - Bacteria: Pfam Domain and Motif
  - Immune Epitope Database (IEDB) Epitope Search (*influenza*)
  - Gene Ontology (GO) Keyword Search (bacteria, Microsporidia)
  - User-interactive blast (blastn, blastp, blastx)
  - Blastp alignments (top 10 + detailed report)
    - SNP details, Amino acid consensuses, IEDB validated epitope data display
  - Additional GBrowse tracks
    - Added IEDB Eptiope track, SNPs, Predicted MHC Class I epitopes (*influenza*)
    - Added SNP track and display for the concatenated reference sequence *A/Puerto Rico/8/34*
    - Added PSIPRED Prediction track (*Francisella Schu 4*)
    - Blastp alignments (top 50), Pfam domains/motifs, Glimmer ORFs, 6-frame translation
- Tools: Added user-interactive multiple sequence alignment (MUSCLE)
- New Features
  - Gene Cart allowing users to store and download sequences and annotations
  - Add linkouts from NCBI Entrez gene ID, taxon ID, genome accession, UniProt accession, and other locations.
  - For flu, dynamically listed the appropriate hosts, subtypes, and genes
  - Added additional influenza consensus sequences for non-human hosts. Dynamically populated the subtype list based on host, and the segment list based on subtype
  - Column sorting for search result tables

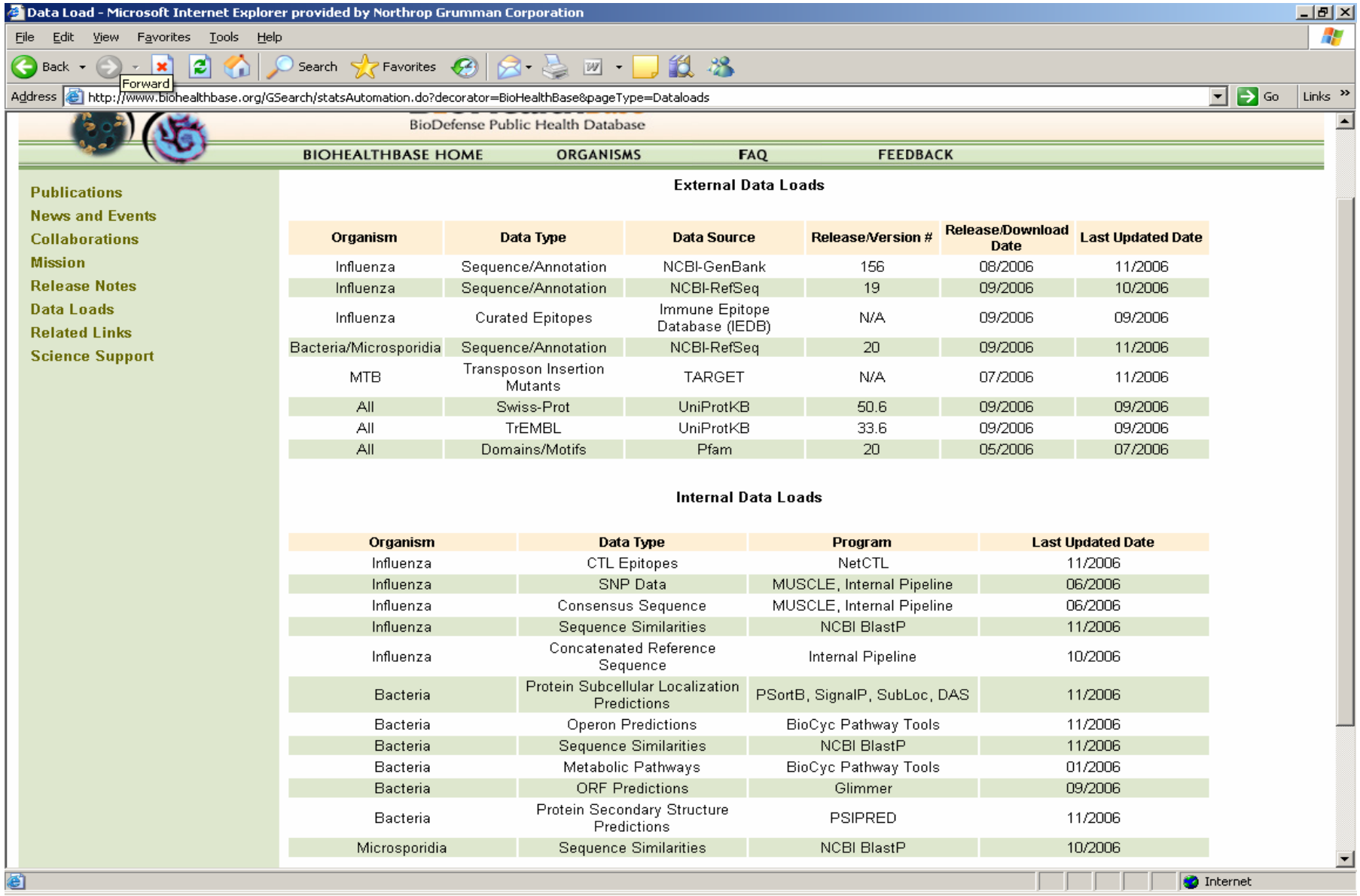
Through V2.2 Release - Data

- New
  - *Influenza* sequences
  - Amino acid consensuses (*influenza*)
  - SNP data and concatenated sequences (*influenza*) *Encephalitozoon cunniculi*
  - *Francisella LVS, Schu 4* (protein secondary structure predictions), *FSC 198* (glimmer predictions/protein subcellular localization)
  - *Mycobacterium sp. MCS* (and glimmer predicions/protein subcellular localization)
  - Immune Epitope Database (IEDB) validated epitopes and NetCTL epitope predictions
- Protein information, integrated with corresponding NCBI annotation and sequences, with updates from:
  - UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, Pfam (domains and motifs)
- Blastp alignments for all BioHealthBase proteins mapped against the SwissProt protein database
- Updates – **Monthly**
  - Internal : October 2006
  - **BRC Central (gff3)**: November 2006

Key BioHealthBase Features

- BioHealthBase is an **integrated** resource
  - Interrelates data from NCBI, UniProt, Pfam, and other sources
    - Direct summary and visualization of integrated data
    - Linkouts to source data sites for additional data details
  - Allows "one-stop shopping" for science researchers
- BioHealthBase provides value-added, **pre-computed** data
  - Valuable processed data not available elsewhere
    - Protein structural and functional predictions
    - Sequence alignment and polymorphism analysis
    - Predicted immunological epitopes
- BioHealthBase provides a forum for displaying **collaborative** data
  - Pathways with Reactome; epitopes with Immune Epitope Database
- BioHealthBase emphasizes **host-pathogen interactions**
  - Pathogen effects on host cellular pathways
  - Immune response to pathogen infection
  - Support of research related to vaccine, therapeutic and diagnostic development

BHB Primary Data Sources



External Data Loads

Organism	Data Type	Data Source	Release/Version #	Release/Download Date	Last Updated Date
Influenza	Sequence/Annotation	NCBI-GenBank	156	08/2006	11/2006
Influenza	Sequence/Annotation	NCBI-RefSeq	19	09/2006	10/2006
Influenza	Curated Epitopes	Immune Epitope Database (IEDB)	N/A	09/2006	09/2006
Bacteria/Microsporidia	Sequence/Annotation	NCBI-RefSeq	20	09/2006	11/2006
MTB	Transposon Insertion Mutants	TARGET	N/A	07/2006	11/2006
All	Swiss-Prot	UniProtKB	50.6	09/2006	09/2006
All	TrEMBL	UniProtKB	33.6	09/2006	09/2006
All	Domains/Motifs	Pfam	20	05/2006	07/2006

Internal Data Loads

Organism	Data Type	Program	Last Updated Date
Influenza	CTL Epitopes	NetCTL	11/2006
Influenza	SNP Data	MUSCLE, Internal Pipeline	06/2006
Influenza	Consensus Sequence	MUSCLE, Internal Pipeline	06/2006
Influenza	Sequence Similarities	NCBI BlastP	11/2006
Influenza	Concatenated Reference Sequence	Internal Pipeline	10/2006
Bacteria	Protein Subcellular Localization Predictions	PSortB, SignalP, SubLoc, DAS	11/2006
Bacteria	Operon Predictions	BioCyc Pathway Tools	11/2006
Bacteria	Sequence Similarities	NCBI BlastP	11/2006
Bacteria	Metabolic Pathways	BioCyc Pathway Tools	01/2006
Bacteria	ORF Predictions	Glimmer	09/2006
Bacteria	Protein Secondary Structure Predictions	PSIPRED	11/2006
Microsporidia	Sequence Similarities	NCBI BlastP	10/2006

Existing Collaborations

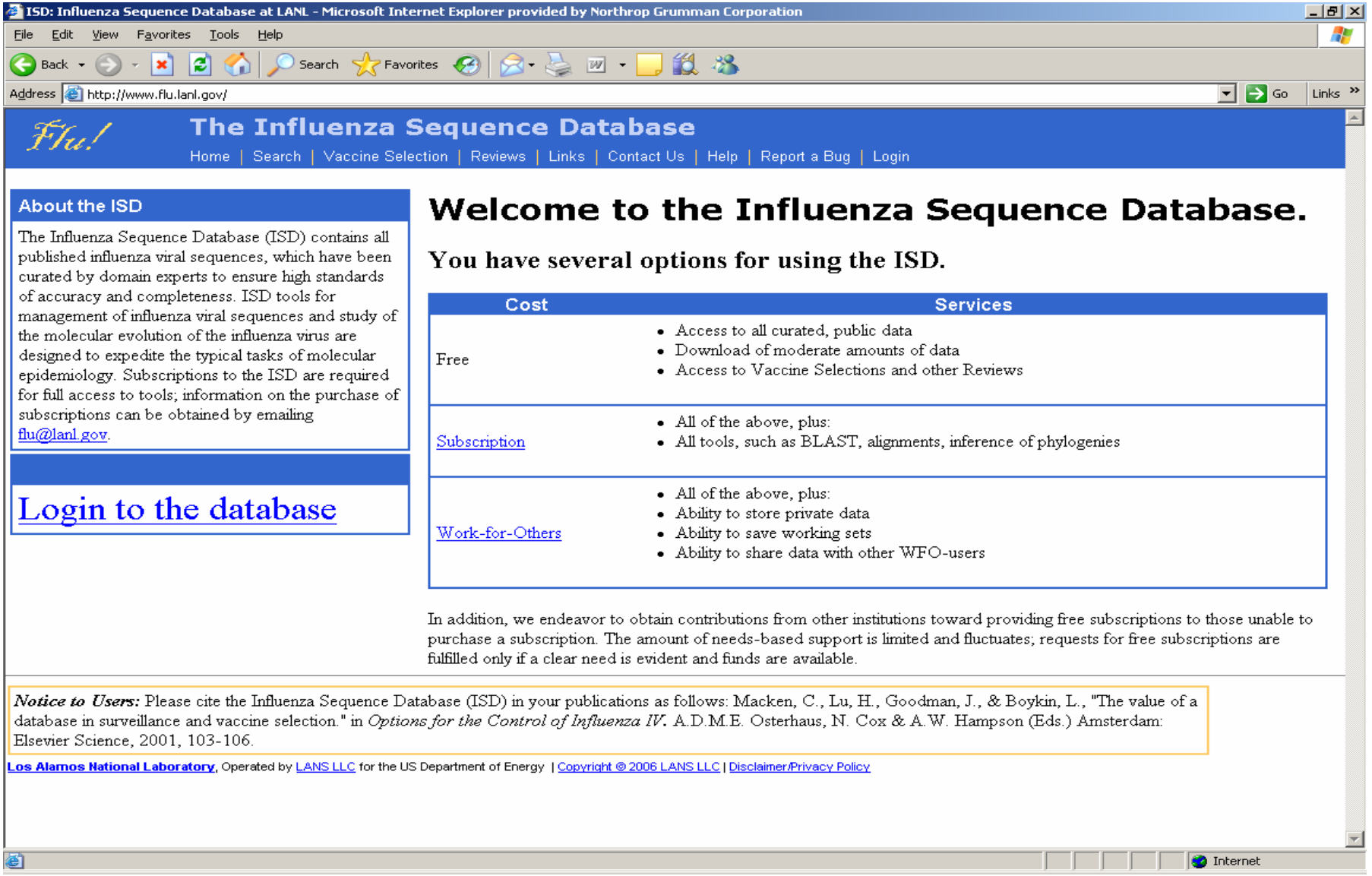
- Reactome - an open source, curated resource of core pathways and reactions in human biology authored by biological researchers with expertise in their fields (**Influenza Infection and Influenza Life Cycle**)

- TARGET (Tuberculosis Animal Research and Gene Evaluation Taskforce) BHB Mycobacterium database provides the visualization of insertion coordinates for each of these mutants using GBrowse as well as a direct link to the Target's and CSU's web-page for ordering the mutant strains.

- BioCyc - integrating metabolic pathway information using the Mycobacterium tuberculosis (Mtbrvcyc) and Francisella tularensis (Frantcyc) pathway databases and using the Pathway Tools and related databases (http://biocyc.org) to obtain and update the metabolic pathway information for BHB pathogens.

- IEDB – For influenza virus, we are integrating their curated database of immune epitope data and analysis resources for antibody and T cell epitope data and the MHC binding data from a variety of different antigenic sources and immune epitope data.

Future Collaboration – Influenza Sequence Database (ISD)



ISD 2-Year Plan

- Interagency Agreement between DOE/LANL and NIH/NIAID
- 2 Year Timeline: January 2007 – December 2008
- Three Phases
- End Result – Integrate LANL ISD content and functions into BHB BRC
- Engage Catherine Macken, one of our SWG members, as a co-PI for the BHB BRC

Future directions

- Current plans
  - Continue to enhance pathway data
  - Phylogenetic tree construction
  - **3D protein** chemical structure **visualization**
  - Additional manual curation
  - Add *Giardia lamblia* and *ricinis communis* into BHB site/resource

- Advice and feedback
  - feedback@biohealthbase.org

3D Protein Visualization

BioHealthBase
BioDefense Public Health Database

- Summary Requirements
  - Implement 3D visualization tool
  - Map sequence features onto 3D visualization views
  - Map multiple sequence features onto 3D visualization views
  - Map non-linear sequence features onto 3D visualization views
  - Map 3D protein view as amino acids
  - Enable search in 3D
  - View sequences feature values
  - Visualize 3D structure of composite sequences from multiple sequences
  - Visualize binding
  - Save 3D visualization as movie or photo (if not built in)
  - Create visual shortcuts for popular views
  - Align primary, secondary and 3D structure in a collapsible single view
  - Integrate protein-to-protein binding database data to be able to bind companion proteins to original selected protein in 3D
  - Enable oligomerization or multiple entity binding of monomers

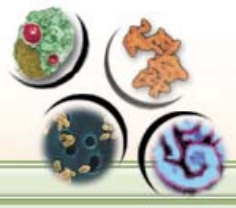
3D Visualization - Functional Requirements

- Display 3D structure of proteins
- Mapping / Highlighting
  - Highlight genome features on 3D structure
  - Highlight multiple genome features
  - Highlight based on annotation characteristic threshold
- Search
  - Search for region based on sequence, annotation
- Output
  - Save publication quality image, movies
- Binding
  - Protein-to-protein
  - Protein-to-small molecules
- Comparison
  - Multiple sequence alignment in 3D
  - Sequence similarity
    - Highlight differences, similarities

3D Visualization Application Evaluation

- Criteria
  - Web application
  - Open source
  - Cross-platform
  - Browser independent
  - Require no installation
  - Native PDB support
  - Easily integratable
  - Allow highlighting or feature mapping

- Evaluations
  - Jmol
  - Protein Explorer (RasMol)
  - Chime
  - FPV
  - AutoDock
  - Chimera
  - Cn3D
  - MICE
  - QuickPDB
  - VMD
  - Strap

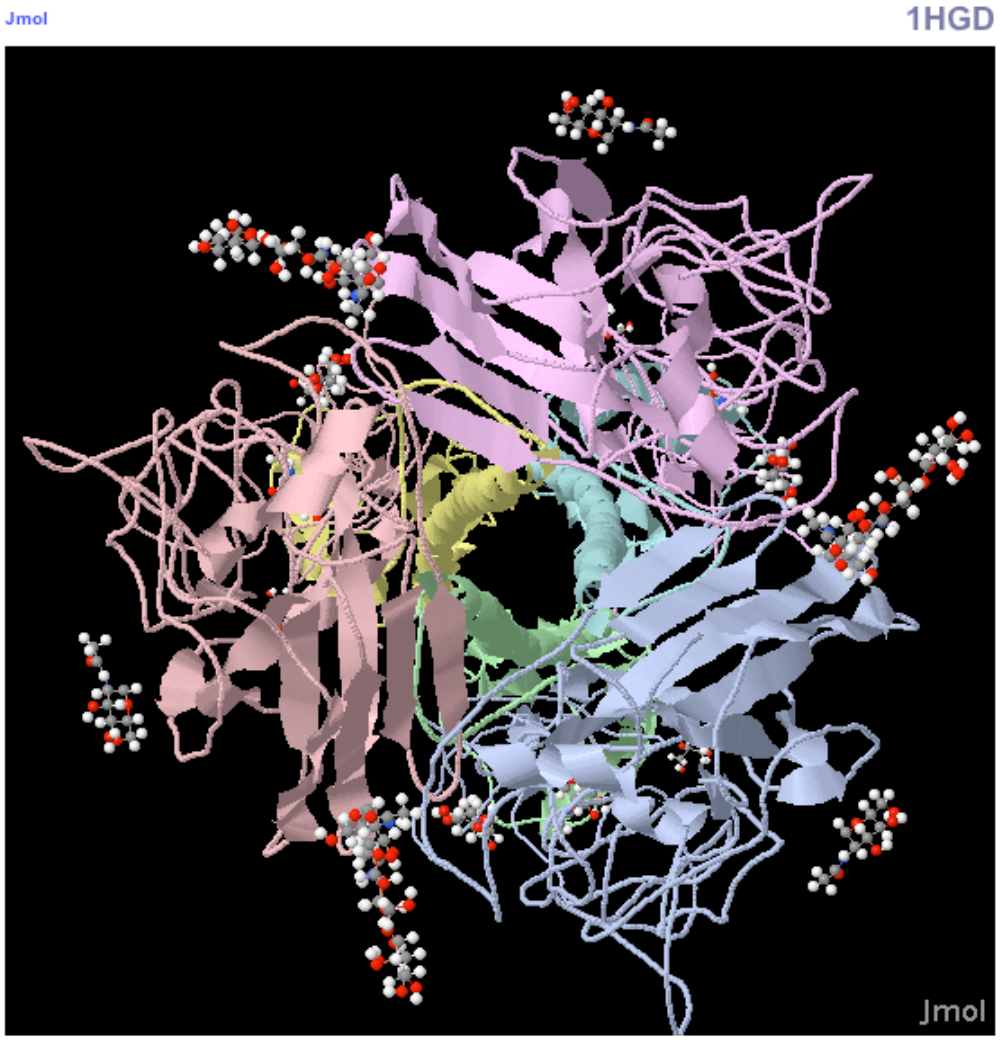


Structural Biology in BHB

BioHealthBase
BioDefense Public Health Database

- 600 structures from four organisms
- Primarily Influenza and Mycobacterium
- 0 structures for Francisella or E cuniculi
- Viewed in Jmol applet
- Processing offloaded to client side

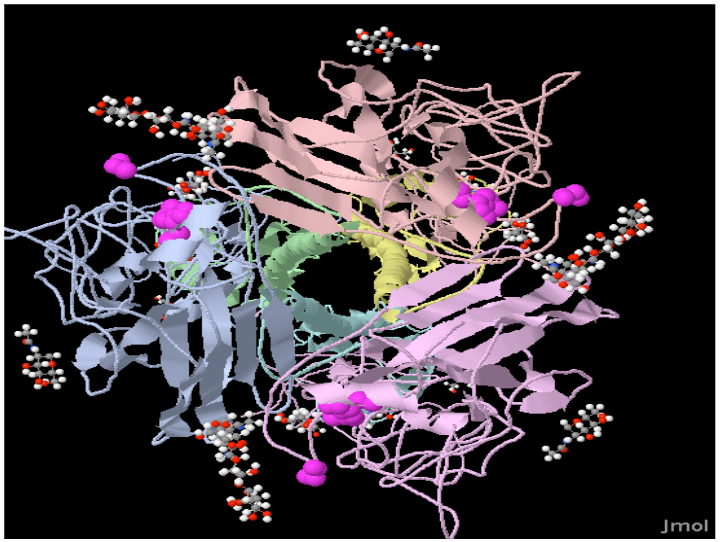
Influenza *HA trimer binding with sialic acid*



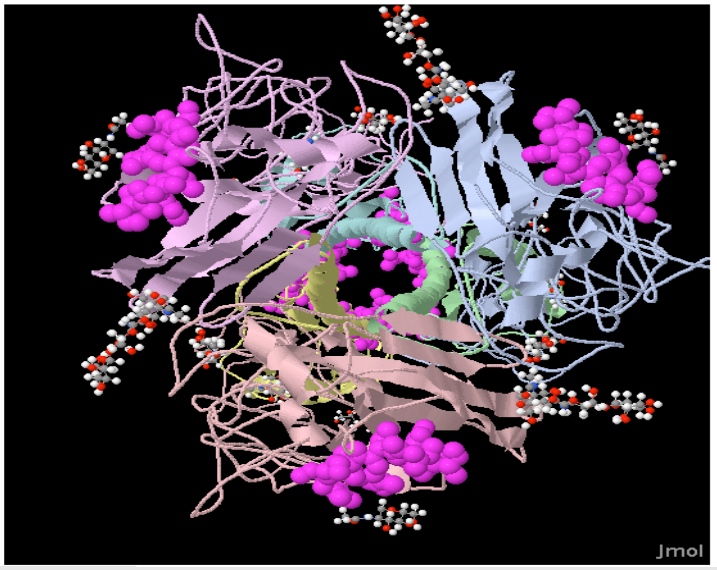
Jmol 1HGD

Jmol 1HGD

Jmol 1HGD



Jmol 1HGD



Progress in BHB visualization

BioHealthBase
BioDefense Public Health Database

- **Present**: (Stock)
- Load all 600 Structures into GUS 3.5
- Enable viewing of those structures

- **Near Term**: (Mods)
- Link to Gbrowse features
- Search for structures
- Annotate features

- **Future**: (Custom Algorithms)
- Substructure searching
- Threading onto consensus sequence

Acknowledgments

- U.T. Southwestern
  - Burke Squires
  - Feng Luo
  - Shubhada Godbole
  - Jennifer Cai
  - Jyoti Shah
  - Jamie Lee
  - Cathy Spranger
  - Victoria Hunt
- SRI
  - Peter Karp
  - John Pick
- Reactome
  - Marc Gillespie
  - Peter D-Eustachio
- Northrop Grumman
  - Kevin Biersack
  - Ed Klem
  - Carey Gire
  - Jason Lucas
  - Sharmila Vattikuti
  - Sanjeev Kumar
  - Paul Shrabstein
  - Surabhi Sharma
  - Tammie Ajayi
  - Aihui Wang
  - Zuoming Deng
  - Jianjun Wang
  - DeWayne Ejikeme
  - Soumya Sengupta
  - Doug Marcey
- Vecna
  - James Wolowicz
  - Chris Larsen
  - Al Ramsey
- Amar
  - X. Wei

Acknowledgments - Continued